NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

FORECASTING ARMY ENLISTED ETS LOSSES

by

Gregory J. Whelan

June 2013

Thesis Advisor: Samuel E. Buttrey
Second Reader: Chad W. Seagren

Approved for public release, distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704–0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202–4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704–0188) Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2013	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE FORECASTING ARMY ENLISTED ETS LOSSES **6. AUTHOR(S)** Gregory J. Whelan	**5. FUNDING NUMBERS**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943–5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. government. IRB Protocol number ______N/A______.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release, distribution is unlimited	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Army currently uses time series models to forecast active-duty enlisted personnel losses. These time series models can provide accurate predictions but offer no insights into the underlying causes of loss behavior. In order to quantify the various forces that influence retention rates, a regression model is necessary. In this thesis, logistic regression is used to estimate end of term-of-service (ETS) losses. The model estimates the probability of reenlistment for soldiers with 12 months remaining on their enlistment contract. The model relies largely on individual soldier information such as pay grade, military occupation, and education, but also examines the impact of the civilian unemployment rate. Two models are developed. The first model includes 14 main effects. The second model includes the same 14 main effects plus 21 highly significant two-way interaction terms. Both models estimate the total number of personnel that reenlist in a seven-month test period fairly well, although the main-effects model results are more accurate. The two-way interaction model performs slightly better on most statistical measures of model effectiveness. Because the two-way interaction model is more complicated to produce, and does not generate results that are clearly better than the main effects model, this thesis recommends using the main effects model to complement the current set of time series models.

14. SUBJECT TERMS Logistic Regression, Personnel, Manpower, Losses, Retention, Forecasting	**15. NUMBER OF PAGES** 69
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

FORECASTING ARMY ENLISTED ETS LOSSES

Gregory J. Whelan
Major, United States Army
B.A., Wake Forest University, 1997

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
June 2013

Author: Gregory J. Whelan

Approved by: Samuel E. Buttrey
Thesis Advisor

Chad W. Seagren
Second Reader

Robert F. Dell
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Army currently uses time series models to forecast active-duty enlisted personnel losses. These time series models can provide accurate predictions but offer no insights into the underlying causes of loss behavior. In order to quantify the various forces that influence retention rates, a regression model is necessary. In this thesis, logistic regression is used to estimate end of term-of-service (ETS) losses. The model estimates the probability of reenlistment for soldiers with 12 months remaining on their enlistment contract. The model relies largely on individual soldier information such as pay grade, military occupation, and education, but also examines the impact of the civilian unemployment rate. Two models are developed. The first model includes 14 main effects. The second model includes the same 14 main effects plus 21 highly significant two-way interaction terms. Both models estimate the total number of personnel that reenlist in a seven-month test period fairly well, although the main-effects model results are more accurate. The two-way interaction model performs slightly better on most statistical measures of model effectiveness. Because the two-way interaction model is more complicated to produce, and does not generate results that are clearly better than the main effects model, this thesis recommends using the main effects model to complement the current set of time series models.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AUC	Area Under the Curve
AFQT	Armed Forces Qualification Test
CBO	Congressional Budget Office
DoD	Department of Defense
ETS	End of Term of Service
MAPE	Mean Absolute Percentage Error
MOS	Military Occupational Specialty
METS	Months to ETS
NPS	Non-Prior Service
ROC	Receiver Operating Characteristic
RMC	Regular Military Compensation
RSS	Residual Sum of Squares
SRB	Selective Reenlistment Bonus

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Army currently uses time series models to forecast active-duty enlisted personnel losses. These time series models can provide accurate predictions but offer no insights into the underlying causes of loss behavior. In order to quantify the various forces that influence retention rates, a regression model is necessary. In this thesis, logistic regression is used to estimate the probability of reenlistment for soldiers with 12 months remaining on their enlistment contract.

The data set used in this thesis was provided by the Army G-1 and consists of 84 monthly snapshots of the entire active component enlisted force over a seven-year period from October 2005 through September 2012. Each snapshot contains more than 400,000 soldier records and nearly 200 fields. Based on a review of attrition and retention literature, 13 fields that are likely to be good predictors of reenlistment decisions are selected. The 13 variables are gender, term of service, pay grade, reenlistment eligibility, Armed Forces Qualification Test (AFQT) score, military occupational specialty (MOS), marital status, number of dependents, number of administrative flags, race, education level, type of accession, and months of active federal service. Two other variables that were not part of the original data set are also examined: the civilian unemployment rate and the consumer confidence index. After evaluating each of the variables one at a time and collectively, all of the variables except consumer confidence index are found to be good predictors. From there two models are developed. The first model includes the 14 main effects. The second model includes the same 14 main effects plus 21 highly significant two-way interaction terms. Both models estimate the total number of personnel that reenlist in a seven-month test period fairly well, although the main effects model results are more accurate. The two-way interaction model performs slightly better on most statistical measures of model effectiveness. Because the two-way interaction model is more complicated to produce, and does not generate results that are clearly better than the main effects model, this thesis recommends using the main effects model to complement the current set of time series models.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank Professor Buttrey and Major Seagren for their assistance with this thesis. They were always willing to meet with me with little or no warning, and spent many hours advising me. I'd like to thank Major Seagren in particular for recommending *Applied Logistic Regression* by David W. Hosmer and Stanley Lemeshow. The book is easy to understand and provides a clear model-building strategy that was very useful. I recommend it to anyone interested in logistic regression. I'd also like to thank MAJ Rob Erdman and the rest of the Army G-1 staff for providing the raw data used in this thesis, answering data-related questions, periodically reviewing my work and offering constructive criticism. Last but certainly not least, I would like to thank my wife, Jerelyn, for being so supportive during my time here at NPS.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

The Army continually manages its personnel in order to both maximize alignment with force structure requirements, and maintain end strength at congressionally mandated levels. Accurate loss estimates are particularly important because they allow the Army to adjust recruiting, retention, and promotion policies in order to shape the force more effectively. Overestimating Army-wide losses could result in an oversized force that must cut funding from training or research budgets in order to pay the extra personnel, while underestimating losses could result in an undersized force that is unable to fill critical positions.

The Army G-1 staff uses several different time series techniques, such as weighted averages, exponential smoothing, and auto-regression, to forecast personnel losses. Time series techniques are particularly powerful when the process of interest is stationary. Stationarity simply means that the distribution of the outcomes are constant over time. When a process is stationary, the future will look a lot like the past. But this is often not the case with personnel losses because important factors regarding soldier retention can change. For example, the availability of jobs in the civilian economy, the intensity and duration of military conflicts, and the size of reenlistment bonuses being offered can all change over time. Although time series techniques can be designed to react to new trends and seasonal factors, they are reactive in nature. New trends must manifest themselves for a while before a time series model can adjust. Also, the current set of time series models provide no information about why the retention process is behaving in a certain way. In order to adjust manpower policies effectively, decision makers must understand the factors that influence retention rates. For instance, how will a small increase in reenlistment bonuses affect retention rates? Are retention rates likely to change once the war in Afghanistan ends? Which soldiers are the most sensitive to these changes? The answers to these types of questions require a different approach.

B. PURPOSE, SCOPE, AND METHODOLOGY

The purpose of this study is to determine whether multivariate regression modeling can be used to forecast Army active component enlisted retention rates. In particular, the study utilizes logistic regression to estimate the probability of soldier reenlistment based on information available 12 months prior to the end of their term of service (ETS). Logistic regression is appropriate when the response variable has only two outcomes. In this case, soldiers are either retained in the Army or become a loss once they reach their ETS date. For modeling purposes, soldiers that extend their current contract or reenlist under a new contract are considered retained. Soldiers that become a loss at the end of their contract, and soldiers that are involuntarily extended by the Army stop-loss policy, are considered losses. The 12-month mark is significant because historically most soldiers could not reenlist until they were within 12 months of their ETS date. In order to make reenlistment predictions for soldiers who are closer to their ETS date, additional models would need to be developed, but 12 months is a good starting point to see if the concept is useful. The study primarily assesses the statistical significance of individual soldier characteristics since personnel data is already maintained by the Army and is readily available. Once suitable models have been developed, they can be applied to soldiers in the current inventory that are within their reenlistment window. If these results are combined with an attrition model for soldiers that are not in their reenlistment window, a retention probability can be determined for every soldier in the inventory for any month within the next year. Then the results can be aggregated in whatever manner the Army would like. For instance, the Army could easily determine the expected number of losses next September by rank and military occupational specialty (MOS) using a pivot table.

C. ORGANIZATION OF THE STUDY

Chapter II describes attrition and retention literature that is relevant to this study. The literature has been broken into two groups. The first group consists of studies that identify potential covariates for the model. The second group focuses on the analytic techniques used to model retention. Chapter III describes the data set and the model building process. Chapter IV assesses the fit of the model through summary measures and cross validation. Chapter V will summarize the important insights and make recommendations for follow-up research.

THIS PAGE INTENTIONALLY LEFT BLANK

II. LITERATURE REVIEW

A. POTENTIAL COVARIATES

1. Factors Affecting Future Levels of Military Personnel

a. Education and Aptitude Testing

The Congressional Budget Office (CBO) cites two previous studies to support the claim that “recruits that are better educated or who score higher on aptitude tests are more likely to…stay in the Army” [1, p. 6]. Because of this conclusion, the Department of Defense (DoD) has two major goals for the quality of its recruits. First, at least 90 percent of non-prior service (NPS) recruits should be high school graduates. Second, at least 60 percent of NPS recruits should score at or above the 50th percentile on the Armed Forces Qualification Test (AFQT) [1, p. 6].

b. Pay and Reenlistment Bonuses

In 2000, the Congress authorized that annual increases in basic pay be 0.5 percent more than the increase in civilian wages, and also increased housing allowances and other pay significantly [1, pp. 12-13]. This was designed to close a perceived pay gap between military personnel and civilian employees. Between 2001 and 2005, the average regular military compensation (RMC) for the entire active duty enlisted force increased almost 14 percent, adjusted for inflation [1, p. 13]. This was significantly higher than comparable civilian wages and leads the CBO to conclude that the increase in compensation should increase retention of first-term active-duty personnel by 25 percent [1, p. 13].

In 2005, the active Army began a serious effort to improve retention rates. In April 2005 it extended its reenlistment window for soldiers from 12 months to 24 months prior to the expiration of their contract [1, p. 10]. The Army also increased the Deployed Selective Reenlistment Bonus (SRB) from $5,000 to $15,000. As a result of these two policies, the active Army spent more on SRBs in 2005 than it spent in the previous four years combined [1, p. 14]. The CBO also notes that SRBs are unnecessary for soldiers with 11 or more years of service because of the lure of retirement pay, which is available after 20 years of service [1, p. 23].

c. Military Occupation and the Civilian Economy

The CBO identifies eight MOSs that may have been subject to increasing civilian competition associated with the ongoing conflicts in Iraq and Afghanistan. In general they find that the Army is able to successfully overcome the increase in civilian competition by increasing SRBs [1, pp. 22–25].

d. Deployments

The CBO examines a variety of research regarding the effect of deployments and finds that the effect of hostile deployments prior to September 11, 2001, on retention is generally positive. In the post-September 11 era, the effect is mixed. In some cases the deployments had no effect on retention, while in other cases they were associated with lower retention. Some of the factors that negatively impacted retention included the stress associated with long work hours both prior to and during deployments, uncertainties surrounding deployment dates, short-notice deployments, insufficient downtime between deployments, and family separation [1, pp. 27–29].

e. Miscellaneous

The CBO notes that educational benefits such as the Montgomery GI Bill, which are designed to improve recruiting efforts, may also discourage reenlistments. This is because soldiers must get out of the Army in order to utilize their benefit by attending college full-time [1, p. 20].

The CBO briefly mentions several other factors that affect retention such as promotion opportunities, job conditions, and time away from home, but does not examine these effects in further detail [1, p. 22].

2. The Effect of Deployments on Service Members

Based on results of focus groups and surveys, the authors of [2] reach many of the same conclusions as the CBO regarding the effect of deployments on retention. Family separation, long work hours, and uncertainty about deployment dates and duties are all considered negative aspects of deployments. On a positive note, they find that unit cohesion and deployment pay can offset some of the negative aspects of deployments. They also find that service members appreciate the opportunity to participate in meaningful missions.

3. Expectations About Civilian Employment Opportunities

The authors of [3] argue that retention depends on service members' expectations about military and civilian compensation, not on actual compensation [3, p. 12]. Unfortunately, they also find that Army officers are particularly prone to overestimating the ease of finding and keeping civilian employment, and underestimate the costs of many military benefits such as healthcare [3, p. 10]. They also point out that gender and race pay differentials are more common in the private sector, which can result in higher retention of female and black officers [3, pp. 39-42].

4. Cash Incentives for Reenlistment

The authors of [4] state that reenlistment rates were fairly stable between 1996 and 2007. Then in 2006 and 2007 the increased number of long deployments caused the effect of deployments on retention, which had been positive, to become negative. The Army responded by increasing SRBs, and by 2009 the economic recession also began to improve retention [4, p. xiii]. The authors state that "bonuses were a critical tool for the Army in meeting its retention objectives in FY 2007 [4, p. xvii]." They also point out that bonuses are more cost-effective than across-the-board pay increases because the amount can vary, and they can be targeted to occupations with shortages [4, p. xxi].

5. What Soldiers Say About Career Continuance

Based on interviews and focus groups conducted in fiscal years 2006 and 2007 (FY06 & FY07), the authors of [5] identify numerous factors affecting both attrition and retention in the Army. Attrition in this context refers to soldiers that fail to complete their contractual term of service, while retention refers to soldier decisions about reenlisting or extending their contract. The factors affecting retention are similar to the results obtained from the other studies, and have already been described. Some of the factors affecting attrition, however, are different, and could also be relevant to this study because soldiers may attrite during their reenlistment window before reaching their ETS date. The authors find that the factors influencing attrition include mental stability, misconduct, adjustment to Army life, and family related issues such as being a single parent.

6. The Effect of Military Pay

The authors of [6] argue that the time is right for DoD to begin reducing military pay increases for at least three reasons. First, the recruiting and retention climate is excellent in part because of high unemployment rates and rapidly rising college tuition costs. Second, DoD is planning to reduce the size of the force, which will reduce recruiting and retention goals. Third, due to steady military pay increases between 2000 and 2010, and stagnant civilian wages during the same period, Army median RMC grew from the 60th percentile to the 80th percentile of comparable civilian wages (see Figure 1).

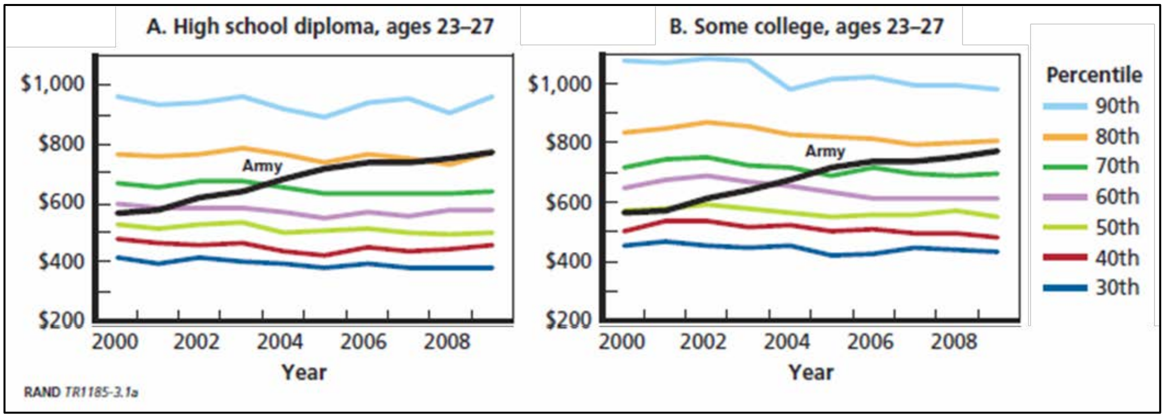


Figure 1. Real Civilian Wages and Median RMC, 2000–2009, in 2010 Dollars. From [6, Fig. 3.1].

During this same time period, civilian health plan premiums increased approximately 150 percent—a growth rate that far exceeded the 31 percent increase in the cost of living [6, p. xiii]. The authors believe that the combination of improved pay and increasingly valuable healthcare benefits should keep retention rates high for the foreseeable future.

7. Summary of Potential Covariates

From the CBO report [1] it appears that education, AFQT score, and Montgomery GI Bill status may be important predictors. Number of deployments is also worth examining but there may be interaction effects with family oriented variables such as marital status and number of dependents, as well as financial variables such as pay and bonuses. MOS may interact with deployments since soldiers with combat MOSs are likely to have a more stressful and dangerous experience than support MOSs. MOS is also worth examining in its own right since some occupations are experiencing more competition from civilian employers than others.

From the Army officer study [3] it appears that race and gender should be examined. Also, a variable which measures perceptions about the civilian economy, such as the consumer confidence index, may be as important or even more important than actual measures of the economy like the unemployment rate.

The cash incentives study [4] makes it clear that the size and availability of reenlistment bonuses are an important factor. The study also claims that the relationship between deployments and reenlistments is nonlinear, and reinforces the importance of economic variables such as the unemployment rate.

The career continuance study [5] shows that many of the factors that influence attrition are not the same as the factors that influence retention. In particular, mental stability, misconduct, and family issues could be important predictors.

Finally, the military pay study [6] clearly illustrates the gains in overall compensation that Army personnel have made relative to civilian employees over the last decade. A variable that could quantify the pay gap between military and civilian personnel would be ideal.

B. POTENTIAL ANALYTIC METHODS

1. A Loss Model Built for the Air Force

The authors of [7] develop several models in order to make monthly loss projections for the entire enlisted force. They build attrition and ETS loss models for first-term, second-term, and career-term personnel, as well as models for personnel who become losses during an extension period, and for retirement losses. The authors examine four approaches to time series modeling: constant rate, regression, autoregressive, and straight line running average [7, pp. 7–11].

a. Constant Rate Models

In a constant rate model the observed differences from month to month are the result of random, uncorrelated disturbances to the system, or noise, and the model has the following form:

$$r(t) = c + e(t)$$

where $r(t)$ is the attrition rate at time *t*,

$e(t)$ is the error at time *t*, and

c is a constant (the mean of the time series).

b. Regression Models

A regression model is appropriate when the data exhibits a dependence on another set of variables. For instance, the attrition rate at time t, $r(t)$, might depend on an airman's salary at time t, $s(t)$. This simple linear regression model has the form

$$r(t) = c + (a)s(t) + e(t)$$

where c is the intercept,

a is the slope of the linear relationship, and

$e(t)$ is the normally independently distributed error term with mean zero.

c. Autoregressive Models

An autoregressive model is useful when the output depends on its own previous values. For instance, in a first order autoregressive model the next observation depends only on the last observation, and has the form

$$r(t) = c + (a)r(t-1) + e(t)$$

where c is a constant term that incorporates the mean of the time series and the autoregressive coefficients, and $e(t)$ is the normally independently distributed error term with mean zero.

d. Straight Line Running Average Models

A straight line running average model uses the average of the most recent k months to predict the attrition rate. For example, a 12-month straight line running average is found by summing the previous 12 months of losses and dividing that total by the sum of the previous 12 start-of-month inventories. The model can also incorporate a seasonal adjustment factor for each calendar month. This type of model has the following form:

$$r(t) = [s(t)][x(t)]$$

where $r(t)$ is the attrition rate at time *t*,

$s(t)$ is the seasonal adjustment factor for calendar month *t*, and

$x(t)$ is the 12-month straight line running average.

The authors of [7] used a regression model to project first term ETS losses [7, p. 31]. ETS cohorts were formed containing all the airmen in the Air Force 12 months prior to their ETS date. They used such variables as the fraction of the cohort lost before the ETS year, and the fraction of the cohort that extended. The airmen were also partitioned into groups based on education, term of enlistment (i.e., the number of years of enlisted obligation), and pay grade. Second-term and career-term ETS losses were projected using a variety of autoregressive and constant rate models [7, pp. 41, 46, 50].

2. A Reenlistment Model Built for the Navy

Nelson [8] identifies several shortcomings with the current regression models used by the Navy to predict reenlistment rates, then makes some recommendations for improving the models. The Navy currently uses three separate models based on time-in-service. Zone A includes sailors with 17 months to 6 years of service, Zone B includes sailors between 6 and 10 years of service, and Zone C includes sailors between 10 and 14 years of service. The variables for the models include end strength, unemployment rate, and attrition rate over the past 11 to 15 years. The model shortcomings include violations of mathematical assumptions, inclusion of insignificant variables, and inclusion of variables that are themselves predictions. Because the Navy performs regression on time series data, Nelson checks to see if the data is uncorrelated and finds some instances where the residuals are not independent. He also finds that the unemployment rate is not a statistically significant predictor and should be removed from the model.

3. A Loss Model Built for the Marine Corps

Orrick [9] uses logistic regression to identify Marines that are likely to attrite prior to their ETS date. He initially evaluates the following variables: AFQT score, number of dependents, years of service, age at enlistment, accession location by Marine Corps district, education level, marital status, separation code, contract length, race, and a binary variable indicating if the Marine has combat experience. For each variable, a baseline level is selected and then dummy variables are created for each of the remaining levels. Orrick finds that nearly all of the variables are statistically significant. He then uses data splitting to validate the model. Orrick claims that the model was able to correctly classify Marines who would attrite over 76 percent of the time.

4. Summary of Analytic Methods

The first two studies, [7] and [8] , make it clear that different types of models may be appropriate for different segments of the military population. In particular, attrition and retention should be modeled separately, and relatively new service members are typically influenced by different factors than experienced service members. The third study [9] confirms that logistic regression can be used to successfully predict attrition, and many of the covariates used in the study were similar to the ones already identified in previous literature.

C. SUMMARY OF LITERATURE REVIEW

Most of the literature reviewed in this thesis focuses on attrition rather than retention, and examines either the DoD as a whole or focuses on military services other than the Army. There was even one study that examines Army officers, but not enlisted personnel [3]. By focusing on Army enlisted retention, this thesis attempts to determine both the shared characteristics and the differences between Army personnel and other service members, enlisted personnel and officers, and attrition and retention.

THIS PAGE INTENTIONALLY LEFT BLANK

III. DATA AND METHODOLOGY

A. DATA SUMMARY

This study utilizes two data sets which were provided by the Army G-1. The first data set is called the "301" file. The 301 file consists of 84 monthly snapshots of the entire active component enlisted force from October 2005 through September 2012 (i.e., a seven-year period from FY06 to FY12). Each snapshot contains more than 400,000 soldier records and nearly 200 fields. The second data set is called the "351" file. The 351 file contains information about monthly personnel actions over the same seven-year period as the 301 file. Personnel actions recorded in the 351 file include gains, losses, reenlistments, extensions, promotions, and demotions. Each record also includes an identification number so that personnel actions can be linked to individual data in the 301 file.

Since this study focuses on estimating the probability of reenlistment for soldiers who are 12 months from their ETS date, the 301 file was filtered to include only these soldiers. Each record was then joined to loss and reenlistment records in the 351 file with matching identification numbers in the next 14 months. A 14-month window was selected in order to include personnel actions that were not recorded until shortly after the ETS month. In a small number of cases, individual soldiers had more than one personnel action during their reenlistment window. For instance, a soldier might extend early in the window and then reenlist or become a loss later in the window. In these cases, reenlistment and loss data were used to determine whether a soldier was retained or lost, and extension and stop-loss information was ignored. Extension and stop-loss information was used only when there was no record of reenlistment or loss.

B. DESCRIPTIVE STATISTICS

1. Overview of the Data Set

Because of the need to look 14 months into the future in order to determine if a soldier was retained, and difficulties incorporating the first three months of the data set, only 67 monthly snapshots from January 2006 to September 2012 were actually used in this analysis. The total number of soldiers who were 12 months from their ETS date

(METS) during this period was 300,788. Nearly 42 percent of these soldiers eventually reenlisted or extended, and the remaining 58 percent became losses or were held under the stop-loss policy. Figure 2 shows that the retention rate was particularly low in the first few months of the data set, then spiked briefly, and hovered around the average thereafter. The monthly average number of soldiers with 12 METS was 4,489 with a standard deviation of 988. Because of this large amount of monthly variation, it will be important for loss planners to track these fluctuations.

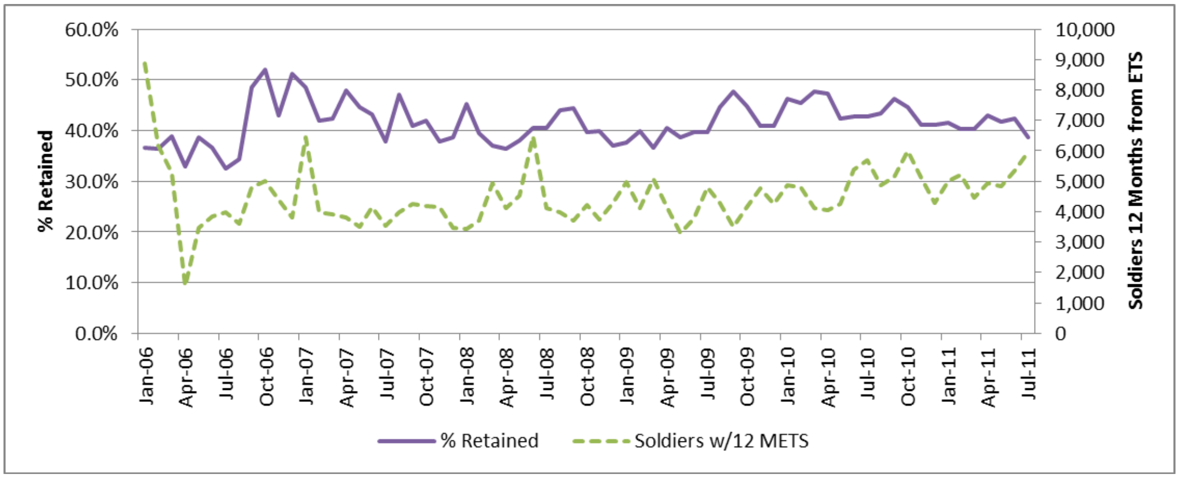


Figure 2. Monthly Number of Soldiers at 12 METS, and Percent of Soldiers That Were Retained

2. Covariates

Although the data set includes nearly 200 variables, many of these were unlikely to have any impact on the decision to reenlist. Some of the others were highly redundant. For instance, there were at least four slightly different ways to measure education level. Some variables contained too many levels to be useful, and some contained too many errors or omissions. Based on these constraints and the information obtained in the literature review, 15 covariates were examined for inclusion in the model. These covariates are examined in greater detail in section D of this chapter.

Unfortunately, the data set does not contain enough information to model the effect of deployments accurately. Although information was available about overseas tours, which include non-combat assignments, it could not be filtered to include only combat deployments. The data set also does not contain enough information about reenlistment bonuses to determine the amount of money soldiers were eligible to receive for reenlisting. This would undoubtedly be a very difficult number to quantify because bonus amounts often depend on many factors such as whether or not the soldier is deployed, time in service, and the length of the reenlistment contract just to name a few. These issues are addressed again in the Follow-up Research Recommendations Section of Chapter V.

C. LOGISTIC REGRESSION

One of the distinguishing characteristics between logistic regression and linear regression is that the response variable in logistic regression is binary [10, p. 1]. Binary variables are typically represented by a value of one if the event of interest occurs, and a value of zero if it does not. Expected values between zero and one can then be interpreted as the probability that the event will occur. If linear regression is applied to a binary response variable it can produce probabilities that are less than zero or greater than one, which are by definition impossible. Instead of a linear relationship between the covariates and the response variable, logistic regression uses a link function known as "log odds" in order to ensure that outcomes stay between zero and one. The odds of an event occurring is defined as the ratio of the probability that an event will occur to the probability that it will not. The conditional mean is expressed as $E(Y \mid x)$ where Y denotes the binary response variable, and x denotes a value of the covariate. If we let $\pi(x)$ represent the conditional mean of Y given x when logistic regression is used, then

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

and the logistic link function is

$$\beta_0 + \beta_1 x = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] \text{ [10, p. 6].}$$

D. COVARIATE DESCRIPTIONS

1. Gender

Over 13 percent of the soldiers in this study are female. Gender was modeled as a nominal variable with 2 levels, male and female.

2. Term of Service

Term of service refers to the number of years of service required by a soldier’s enlistment contract. As shown in Table 1, three-year and four-year contracts are by far the most common. “Blank” and “0” are errors. “Z” represents an indefinite contract. Only senior enlisted personnel are offered an indefinite enlistment. Soldiers with an indefinite term of service get their ETS date extended each time they are promoted. Term of Service was modeled as a nominal variable with seven levels. Some low density levels were grouped together for modeling purposes. See Table 1.

Original Levels

Service Term	Soldiers	% of Total
blank	26	0.0%
0	210	0.1%
1	2,940	1.0%
2	5,990	2.0%
3	129,764	43.1%
4	110,514	36.7%
5	26,033	8.7%
6	18,298	6.1%
7	48	0.0%
8	254	0.1%
9	52	0.0%
Z	6,659	2.2%

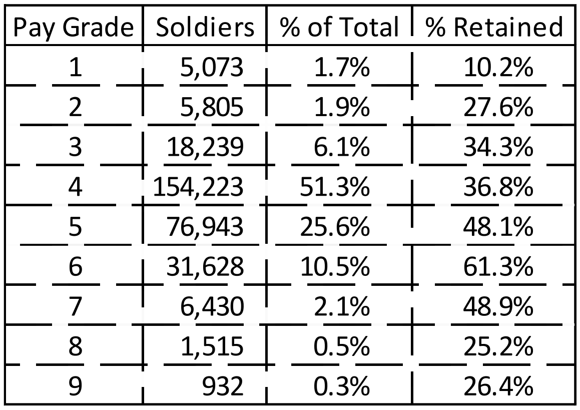
Levels Used in Model

Service Term	Soldiers	% of Total	% Retained
blank, 0-1	3,176	1.1%	26.7%
2	5,990	2.0%	45.8%
3	129,764	43.1%	44.3%
4	110,514	36.7%	41.7%
5	26,033	8.7%	36.8%
6	18,298	6.1%	39.1%
7-9,Z	7,013	2.3%	19.4%

Table 1. Term of Service Distribution

3. Pay Grade

Pay grade is closely related to military rank. E-1 is the lowest enlisted pay grade and E-9 is the highest. E-4 is the most common pay grade in the data set. Pay Grade was modeled as a nominal variable with nine levels.



Pay Grade	Soldiers	% of Total	% Retained
1	5,073	1.7%	10.2%
2	5,805	1.9%	27.6%
3	18,239	6.1%	34.3%
4	154,223	51.3%	36.8%
5	76,943	25.6%	48.1%
6	31,628	10.5%	61.3%
7	6,430	2.1%	48.9%
8	1,515	0.5%	25.2%
9	932	0.3%	26.4%

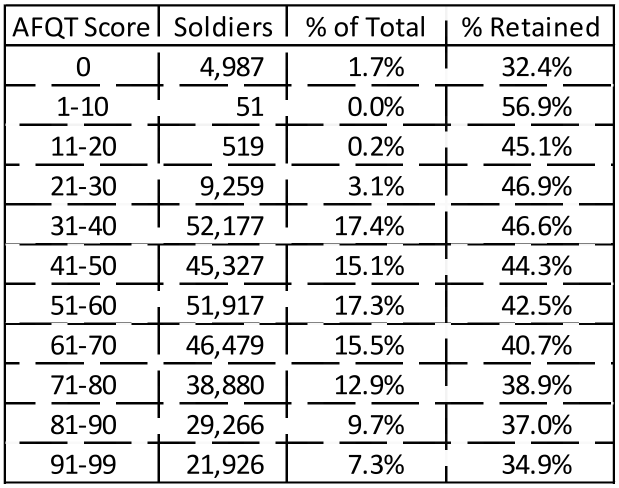
Table 2. Pay Grade Distribution

4. Reenlistment Prohibition Code

Over 16 percent of the soldiers in this study are prohibited from reenlisting for various reasons. See the Appendix for more details. It is important to note that it is possible for soldiers to get their reenlistment prohibition lifted during their reenlistment window. In fact, 29.5 percent of soldiers in the data set who were prohibited from reenlisting with 12 METS eventually reenlisted or extended. Reenlistment Prohibition Code was modeled as a nominal variable with two levels, prohibited from reenlisting and eligible to reenlist.

5. AFQT Score

All prospective recruits take the Armed Forces Qualification Test prior to enlisting in the military. The AFQT score is a percentile with 0 being the lowest score and 99 the highest. AFQT scores can determine whether a potential recruit is allowed to enlist, and the type of MOSs the recruit is able to select. Table 3 shows the distribution of AFQT scores in the data set. The large number of soldiers with a score of “zero” may indicate some type of error in the data set.



AFQT Score	Soldiers	% of Total	% Retained
0	4,987	1.7%	32.4%
1-10	51	0.0%	56.9%
11-20	519	0.2%	45.1%
21-30	9,259	3.1%	46.9%
31-40	52,177	17.4%	46.6%
41-50	45,327	15.1%	44.3%
51-60	51,917	17.3%	42.5%
61-70	46,479	15.5%	40.7%
71-80	38,880	12.9%	38.9%
81-90	29,266	9.7%	37.0%
91-99	21,926	7.3%	34.9%

Table 3. AFQT Score Distribution

AFQT Score could be modeled as a continuous variable or as a nominal variable. If it is modeled as a nominal variable, then there are 100 possible levels, so it will be important to determine how to group the scores into bins in order to reduce the number of levels. Either way it is important to examine how the conditional mean varies as AFQT score changes. Figure 3 indicates that the probability of reenlistment for soldiers with a score of zero is very low. The probability of reenlistment for soldiers with scores between 1 and 37 is much higher, but also highly variable. Finally, for soldiers with scores between 38 and 99 there is a strong linear relationship between AFQT Score and the probability of reenlistment where higher scoring soldiers are less likely to reenlist.

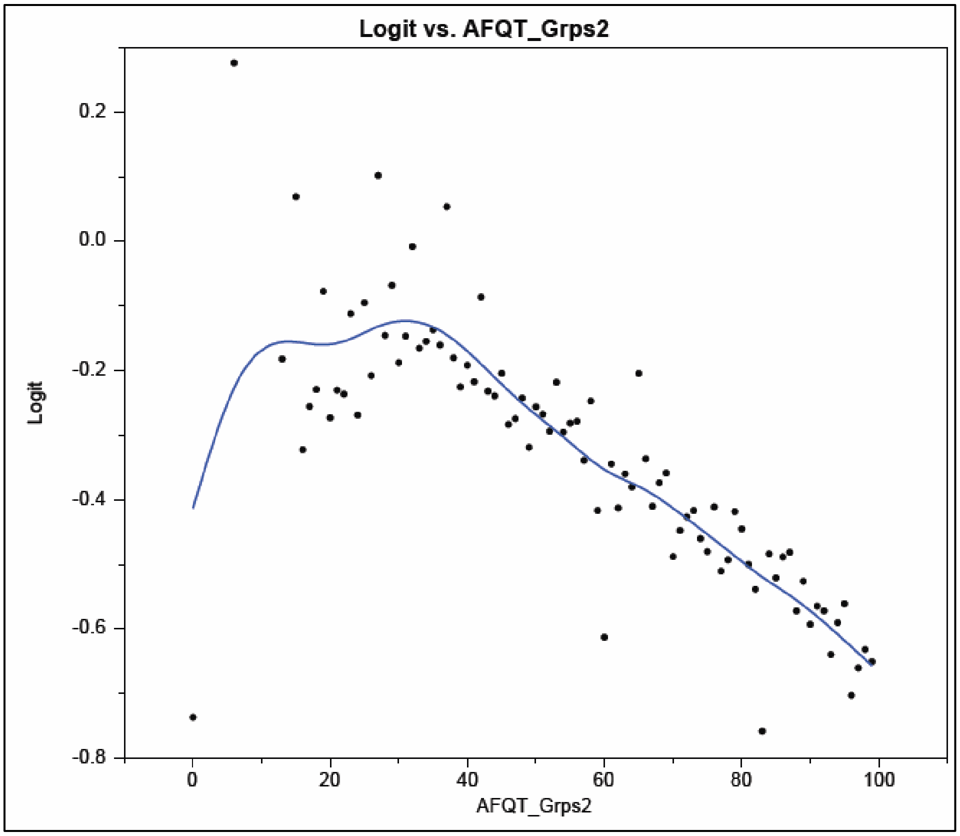
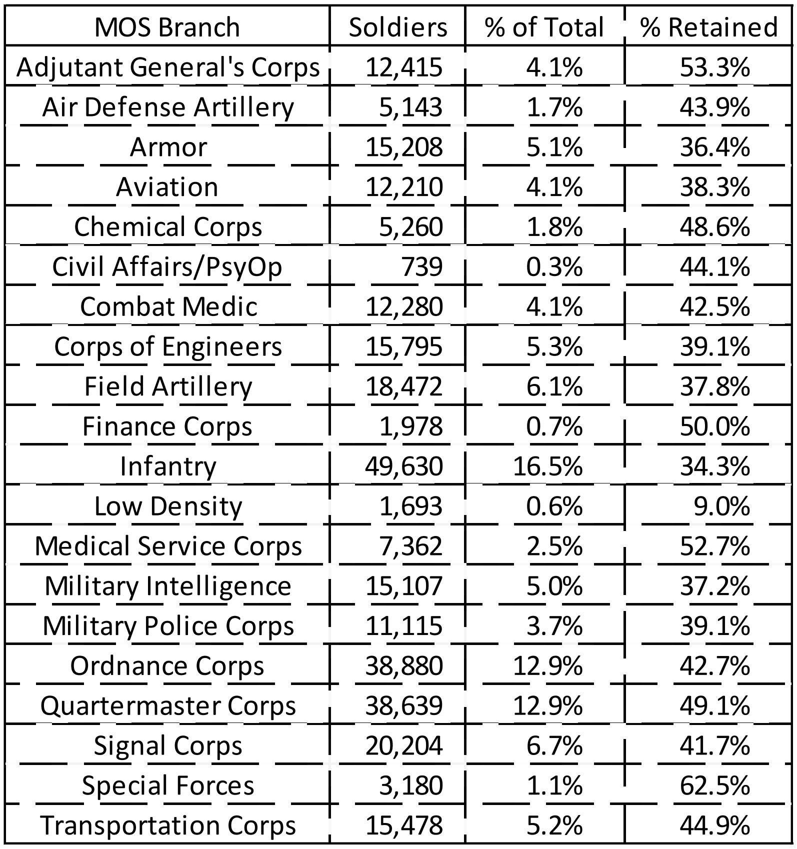


Figure 3. Log-odds of AFQT Scores

Because of the presence of these three distinct groups, AFQT Score was modeled as a piecewise continuous variable. A new nominal variable was added to indicate whether the AFQT score is 0, between 1 and 37, or greater than 37. Then an interaction term was added so that the model uses a different coefficient for each of the three intervals.

6. MOS

There are 292 unique MOSs in the data set. For modeling purposes, the MOSs were grouped into 20 different branches. Table 4 contains a list of the branches and their frequency. MOS Branch was modeled as a nominal variable with 20 levels.



MOS Branch	Soldiers	% of Total	% Retained
Adjutant General's Corps	12,415	4.1%	53.3%
Air Defense Artillery	5,143	1.7%	43.9%
Armor	15,208	5.1%	36.4%
Aviation	12,210	4.1%	38.3%
Chemical Corps	5,260	1.8%	48.6%
Civil Affairs/PsyOp	739	0.3%	44.1%
Combat Medic	12,280	4.1%	42.5%
Corps of Engineers	15,795	5.3%	39.1%
Field Artillery	18,472	6.1%	37.8%
Finance Corps	1,978	0.7%	50.0%
Infantry	49,630	16.5%	34.3%
Low Density	1,693	0.6%	9.0%
Medical Service Corps	7,362	2.5%	52.7%
Military Intelligence	15,107	5.0%	37.2%
Military Police Corps	11,115	3.7%	39.1%
Ordnance Corps	38,880	12.9%	42.7%
Quartermaster Corps	38,639	12.9%	49.1%
Signal Corps	20,204	6.7%	41.7%
Special Forces	3,180	1.1%	62.5%
Transportation Corps	15,478	5.2%	44.9%

Table 4. MOS Branch Distribution

7. Marital Status

Marital Status has nine levels, but more than 99 percent of soldiers belong to one of three groups: Single, Married, or Divorced. Soldiers in the remaining six levels were included in either the Married or Divorced levels as indicated in Table 5.

Original Levels

Marital Status		Soldiers	% of Total
S	Single	140,180	46.6%
M	Married	145,874	48.5%
	Blank	16	0.0%
Z	Unknown	146	0.1%
D	Divorced	13,750	4.6%
A	Annulled	96	0.0%
I	Interlocutory	1	0.0%
L	Legally Separated	534	0.2%
W	Widowed	191	0.1%

Levels Used in Model

Marital Status		Soldiers	% of Total	% Retained
S	Single	140,180	46.6%	34.1%
M	Married	146,020	48.6%	48.3%
D	Divorced	14,588	4.9%	47.7%

Table 5. Marital Status Distribution

8. Number of Dependents

Number of Dependents includes both adult dependents and children.

Dependents	Soldiers	% of Total	% Retained
0	128,048	42.6%	33.7%
1	69,039	23.0%	42.0%
2	47,030	15.6%	47.8%
3	33,136	11.0%	52.0%
4	15,275	5.1%	56.2%
5	5,577	1.9%	57.3%
6	1,860	0.6%	58.1%
7	574	0.2%	61.8%
8	168	0.1%	63.1%
9	52	0.0%	50.0%
10	15	0.0%	53.3%
11	10	0.0%	50.0%
12	3	0.0%	33.3%
13	1	0.0%	0.0%

Table 6. Dependents Distribution

Number of Dependents can be modeled as a continuous or nominal variable. Figure 4 indicates that Number of Dependents increases linearly up through eight dependents, but then declines thereafter. Adding a second order polynomial term to the model incorporates this parabolic shape and provides the best fit with the fewest degrees of freedom. Number of Dependents was modeled as a continuous variable with a second order polynomial term.

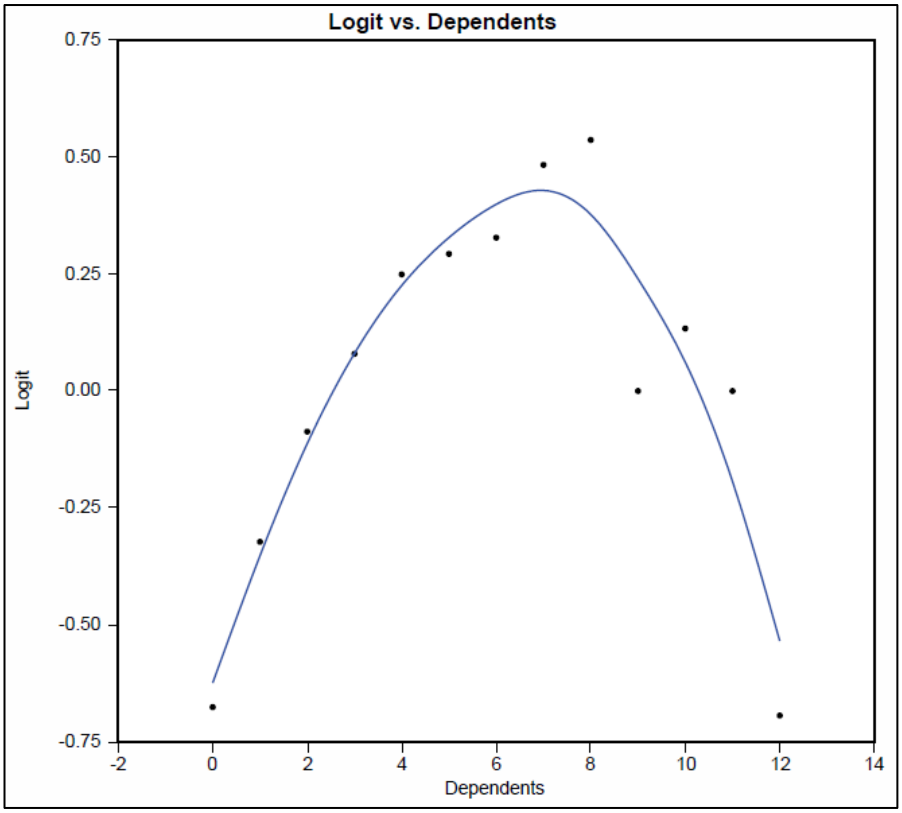
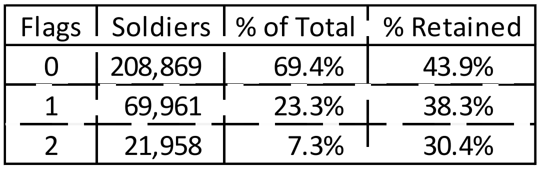


Figure 4. Distribution of Number of Dependents

9. Number of Flags

“Flag” is a term which means suspension of favorable personnel action. Soldiers can be flagged for a variety of reasons including failing a physical fitness test, exceeding body fat standards, misconduct, and many others. The data set was able to record at most two flag codes per soldier. Rather than focusing on the reason for being flagged, the covariate used in the model simply indicates how many times a soldier is flagged. Number of Flags was modeled as a nominal variable with three levels: 0, 1, and 2.

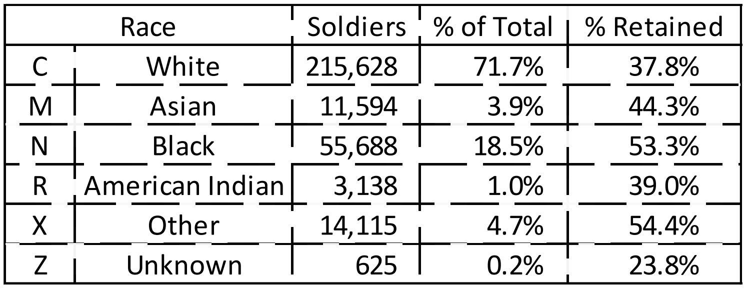


Flags	Soldiers	% of Total	% Retained
0	208,869	69.4%	43.9%
1	69,961	23.3%	38.3%
2	21,958	7.3%	30.4%

Table 7. Distribution of Number of Flags

10. Race

Race was modeled as a nominal variable with six levels as shown in Table 8.

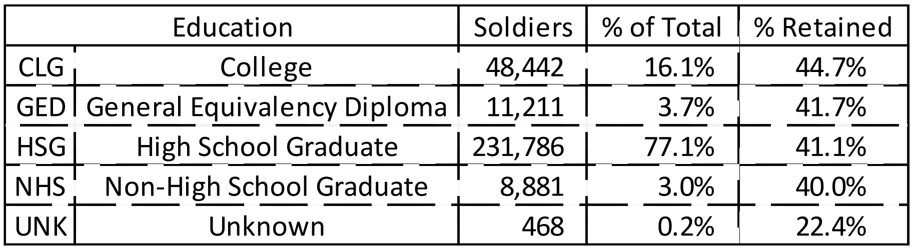


Race		Soldiers	% of Total	% Retained
C	White	215,628	71.7%	37.8%
M	Asian	11,594	3.9%	44.3%
N	Black	55,688	18.5%	53.3%
R	American Indian	3,138	1.0%	39.0%
X	Other	14,115	4.7%	54.4%
Z	Unknown	625	0.2%	23.8%

Table 8. Race Distribution

11. Education Category

Education Category was modeled as a nominal variable with five levels as shown in Table 9.

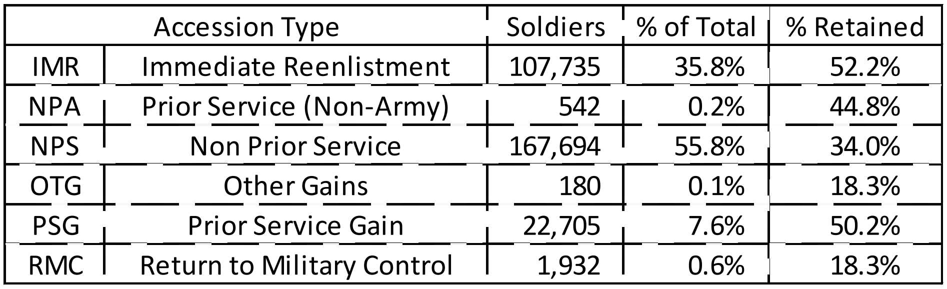


Education		Soldiers	% of Total	% Retained
CLG	College	48,442	16.1%	44.7%
GED	General Equivalency Diploma	11,211	3.7%	41.7%
HSG	High School Graduate	231,786	77.1%	41.1%
NHS	Non-High School Graduate	8,881	3.0%	40.0%
UNK	Unknown	468	0.2%	22.4%

Table 9. Education Distribution

12. Accession Type

Accession Type indicates if a soldier served in the military prior to their current enlistment. The Army defines “prior service” as any applicant with more than 180 days of military service, or those who graduated from military job-training, regardless of time-in-service [11]. As shown in Table 10, a small fraction of soldiers in the study have non-Army prior service, such as Navy or Air Force experience. Most prior service soldiers were in the Army previously, had a break in service, and then decided to get back in the Army. Accession Type was modeled as a nominal variable with six levels.



Accession Type		Soldiers	% of Total	% Retained
IMR	Immediate Reenlistment	107,735	35.8%	52.2%
NPA	Prior Service (Non-Army)	542	0.2%	44.8%
NPS	Non Prior Service	167,694	55.8%	34.0%
OTG	Other Gains	180	0.1%	18.3%
PSG	Prior Service Gain	22,705	7.6%	50.2%
RMC	Return to Military Control	1,932	0.6%	18.3%

Table 10. Accession Type Distribution

13. Months of Active Federal Service (AFS)

Figure 5 shows the distribution of years of AFS in the data set. Since all of the soldiers in the data set are 12 months from their ETS date, and most soldiers sign three-year or four-year contracts, there are very few soldiers with less than two years of service.

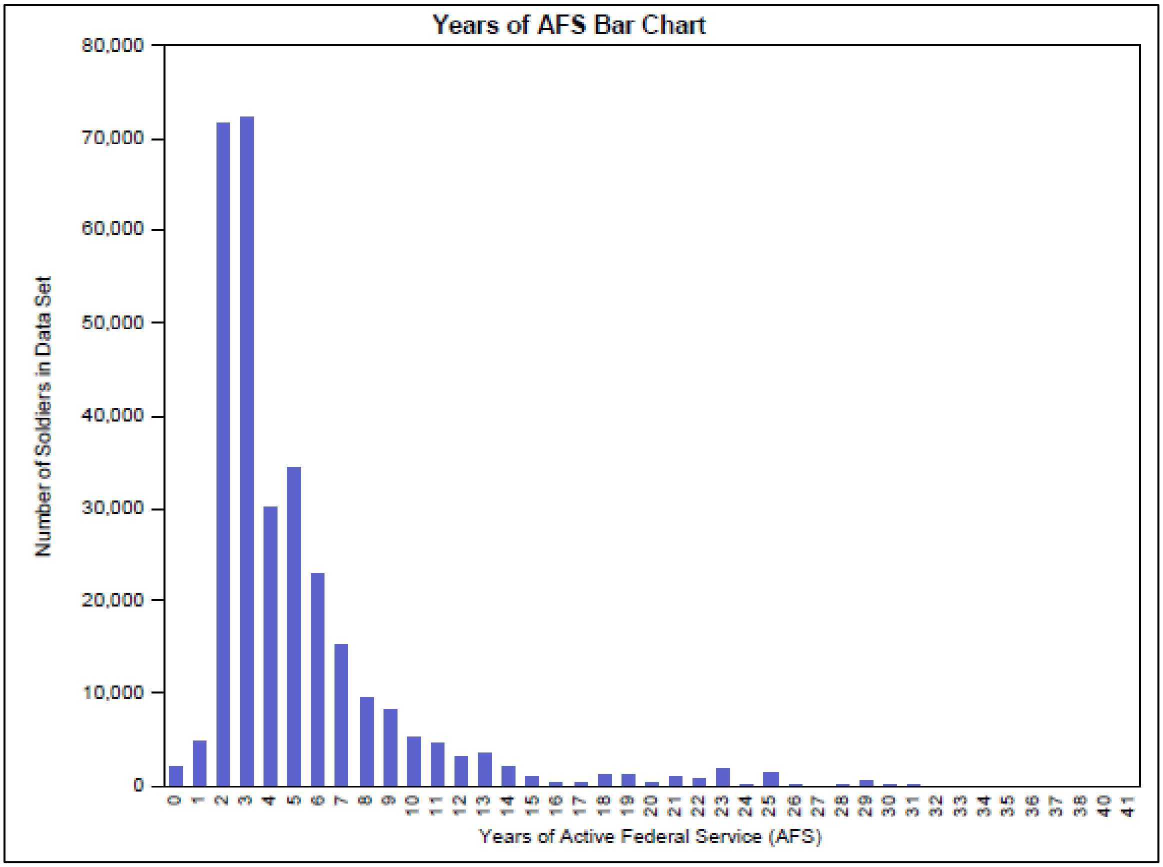


Figure 5. Years of Active Federal Service (AFS) Bar Chart

Just like AFQT Score, Months of AFS could be modeled as a continuous variable or as a nominal variable. Figure 6 shows that the log odds of reenlisting increases linearly as time-in-service increases until soldiers reach 227 months of service, which is one month short of 19 years of service, when it drops significantly. This is clearly related to retirement eligibility, which begins at 20 years of service. As soldiers get closer to retirement eligibility they become increasingly likely to reenlist, but once they qualify for retirement there is much less incentive to remain in the force. It is also true that promotions become more difficult to achieve later in a soldier's career, and soldiers who are not selected for promotion are eventually forced out.

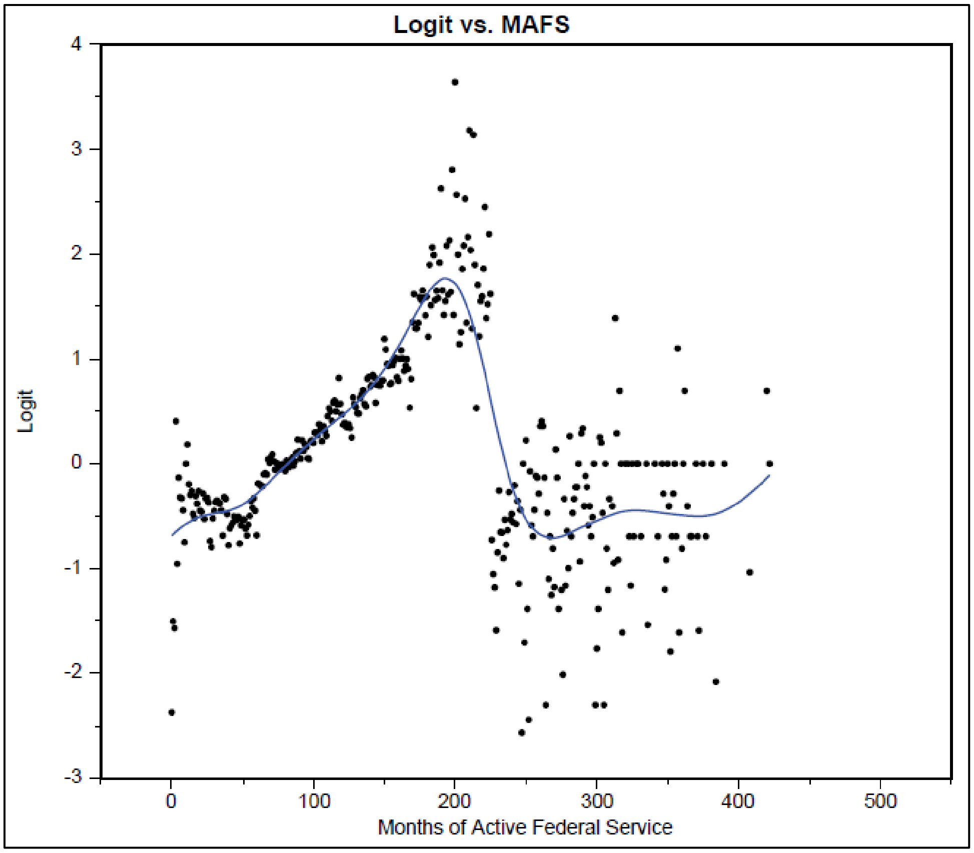


Figure 6. Log-odds of Months of Active Federal Service

Because of these two distinct groups, Months of AFS was modeled as a continuous variable and a second new binary variable was added to indicate whether the duration is less than 227 months, or greater than or equal to 227 months. Then an interaction term was added so that the model uses a different coefficient for the two intervals.

14. Unemployment Rate

The U.S. unemployment rate is obtained by dividing the number of unemployed individuals by the number of individuals in the labor force. As shown in Figure 7, the unemployment rate was less than 5 percent in 2006 and 2007, then rose dramatically in 2008 and 2009, and remained above 9 percent in 2010 and most of 2011 [12].

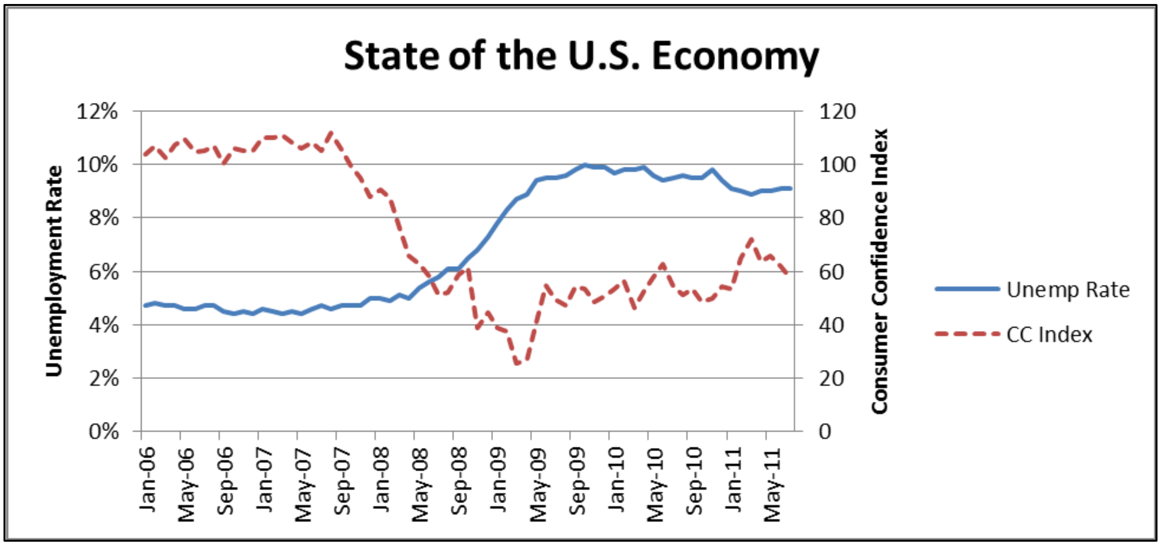


Figure 7. U.S. Unemployment Rate and Consumer Confidence Index, AN 2006–JUL 2011 [12, 13]

Figure 8 shows that the log-odds of reenlisting exhibits a lot of variation as the unemployment rate changes. For some reason the log-odds are occasionally high even when the unemployment rate is low. The only area where there appears to be a linear relationship is when the unemployment rate goes above nine percent. At that point each incremental increase in the unemployment rate results in an increase in the log-odds of reenlisting. One possible explanation for these counterintuitive results during periods of low unemployment is that the Army increased the size and availability of reenlistment bonuses during these months in order to keep retention rates high. It is also possible that most people do not pay attention to the unemployment rate until it becomes very high and begins to get widespread media attention. This is speculation however, and additional research is required in order to adequately explain the results.

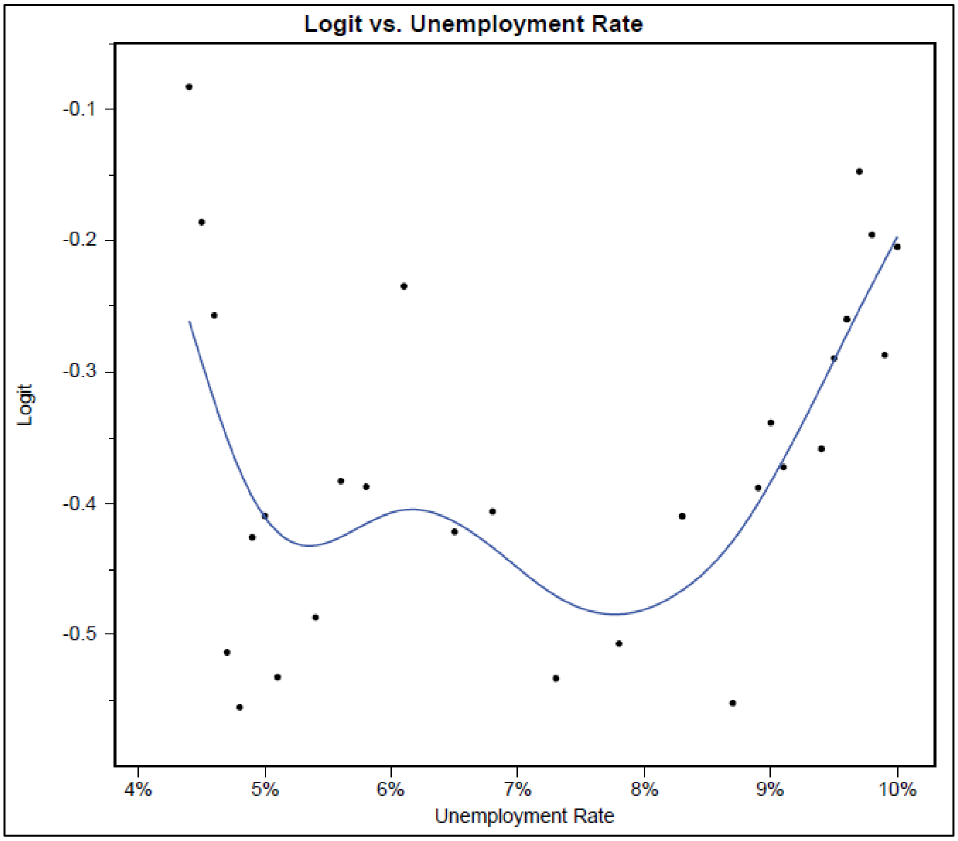


Figure 8. Log-odds of Unemployment Rate

Because of the parabolic shape, adding a second degree polynomial and/or a third degree polynomial to the univariate model produces the best fit. Since the third degree polynomial model is only slightly better than the second degree polynomial model, the third degree polynomial term was not included in order to keep the model as simple as possible and avoid overfitting.

15. Consumer Confidence Index

Consumer Confidence in the United States is reported by The Conference Board. The Conference Board Consumer Confidence Index® (CCI) is a barometer of the health of the U.S. economy from the perspective of the consumer. The index is based on approximately 3,000 completed questionnaires reflecting consumers' perceptions of current business and employment conditions, as well as their expectations for the next six months regarding business conditions, employment, and income [13].

Based on the literature review, CCI seems like a promising covariate. Unfortunately, there is no discernible pattern to the reenlistment log-odds plot in Figure 9. From Figure 7 it is also apparent that the CCI has a strong negative correlation with the unemployment rate, so perhaps it provides no new information anyway. For these two reasons, CCI was not explored any further.

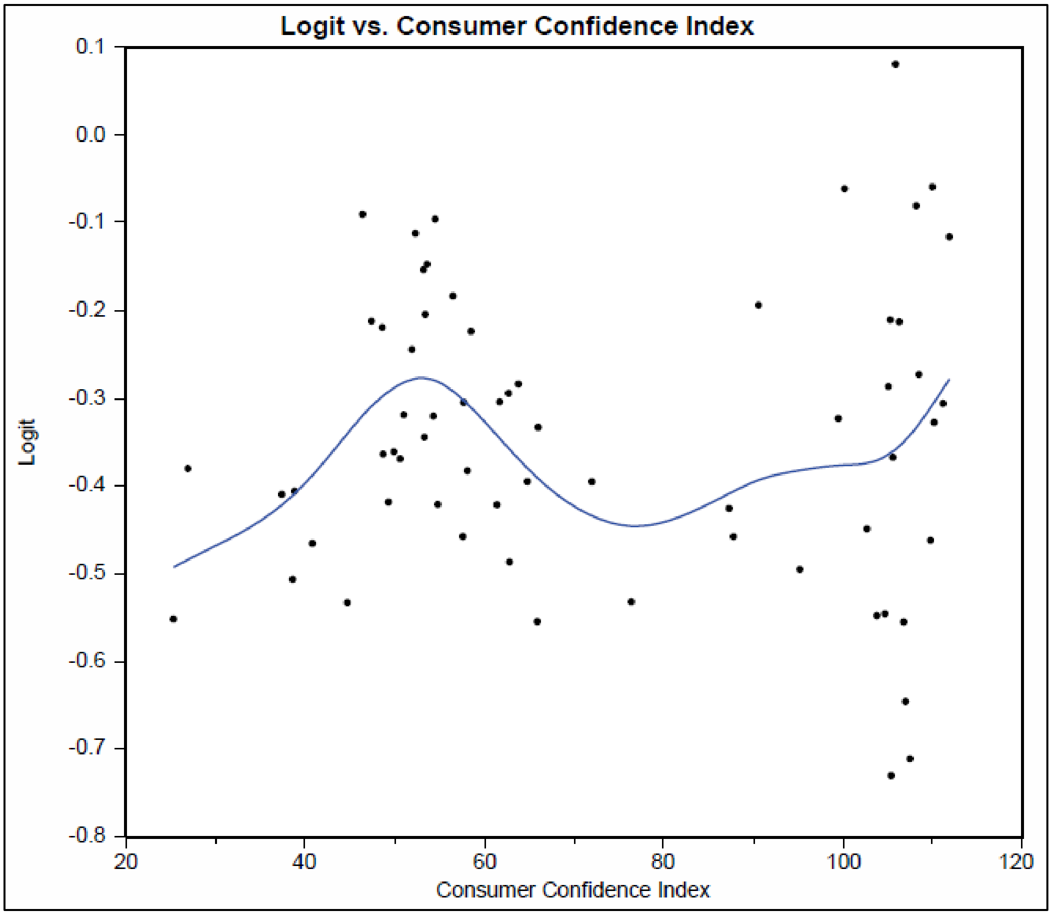
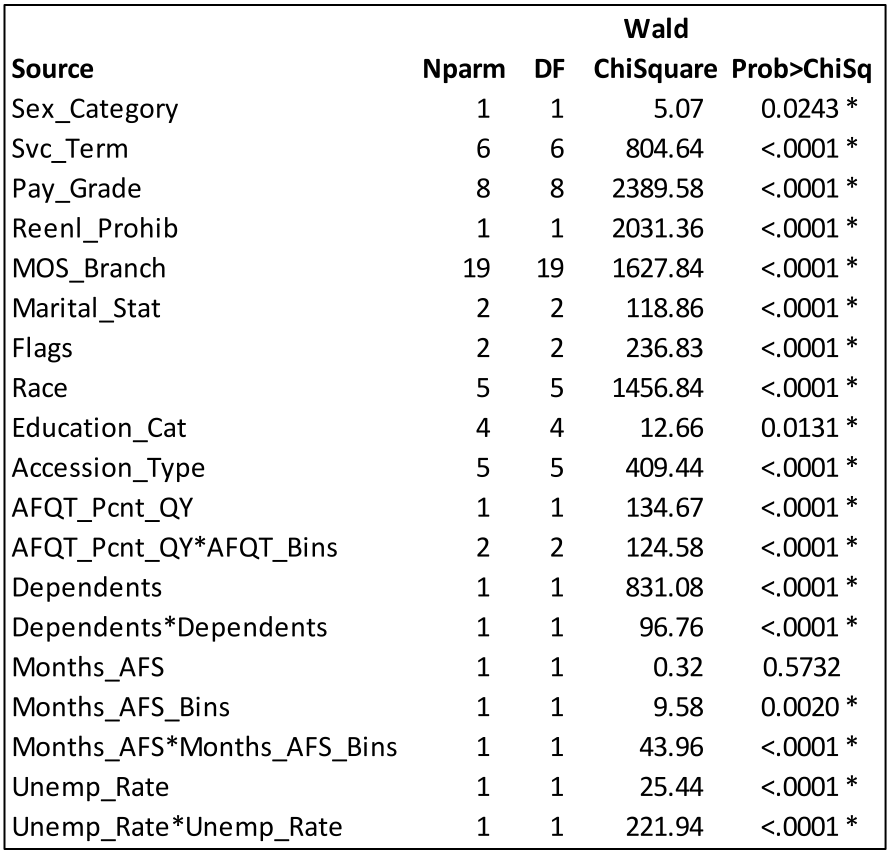


Figure 9. Log-odds of Consumer Confidence Index

E. VARIABLE SELECTION

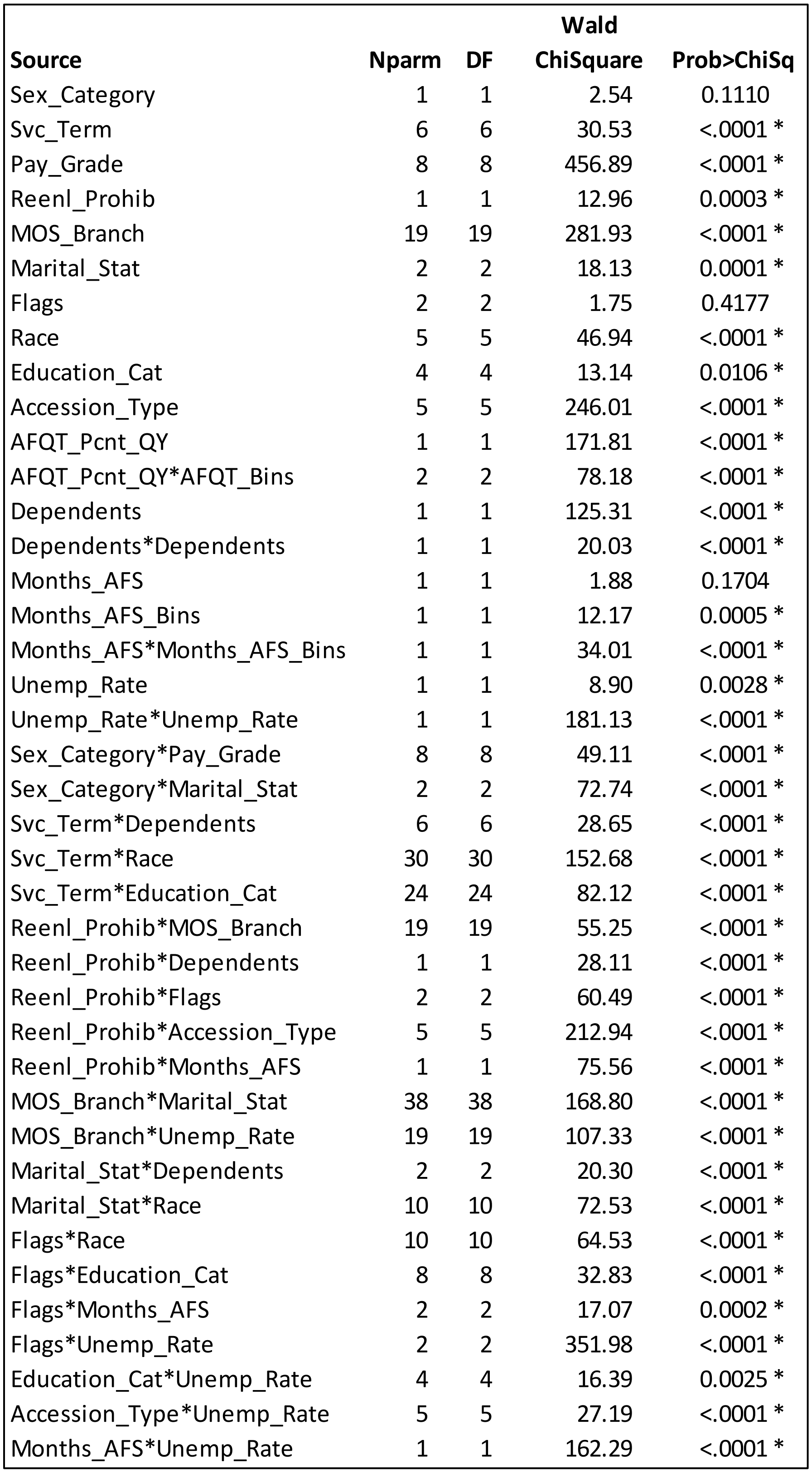
Univariate analysis of the 14 remaining covariates confirms that they are all significant with p-values less than 0.0001. The results when all of the covariates are included in a single model are shown in Table 11. All of the variables are highly significant and should remain in the model.



Source	Nparm	DF	Wald ChiSquare	Prob>ChiSq
Sex_Category	1	1	5.07	0.0243 *
Svc_Term	6	6	804.64	<.0001 *
Pay_Grade	8	8	2389.58	<.0001 *
Reenl_Prohib	1	1	2031.36	<.0001 *
MOS_Branch	19	19	1627.84	<.0001 *
Marital_Stat	2	2	118.86	<.0001 *
Flags	2	2	236.83	<.0001 *
Race	5	5	1456.84	<.0001 *
Education_Cat	4	4	12.66	0.0131 *
Accession_Type	5	5	409.44	<.0001 *
AFQT_Pcnt_QY	1	1	134.67	<.0001 *
AFQT_Pcnt_QY*AFQT_Bins	2	2	124.58	<.0001 *
Dependents	1	1	831.08	<.0001 *
Dependents*Dependents	1	1	96.76	<.0001 *
Months_AFS	1	1	0.32	0.5732
Months_AFS_Bins	1	1	9.58	0.0020 *
Months_AFS*Months_AFS_Bins	1	1	43.96	<.0001 *
Unemp_Rate	1	1	25.44	<.0001 *
Unemp_Rate*Unemp_Rate	1	1	221.94	<.0001 *

Table 11. Wald Tests on Main Effects Model

In order to check for two-way interaction effects, nearly every possible two-way interaction variable was added to the main effects model one at a time. Interaction variables with p-values less than 0.0001 were identified and added to the model collectively. Then the interaction variables were evaluated once again and removed if their p-values were higher than 0.01. This model building process is consistent with the strategy outlined in [10]. The results are shown in Table 12.



			Wald	
Source	**Nparm**	**DF**	**ChiSquare**	**Prob>ChiSq**
Sex_Category	1	1	2.54	0.1110
Svc_Term	6	6	30.53	<.0001 *
Pay_Grade	8	8	456.89	<.0001 *
Reenl_Prohib	1	1	12.96	0.0003 *
MOS_Branch	19	19	281.93	<.0001 *
Marital_Stat	2	2	18.13	0.0001 *
Flags	2	2	1.75	0.4177
Race	5	5	46.94	<.0001 *
Education_Cat	4	4	13.14	0.0106 *
Accession_Type	5	5	246.01	<.0001 *
AFQT_Pcnt_QY	1	1	171.81	<.0001 *
AFQT_Pcnt_QY*AFQT_Bins	2	2	78.18	<.0001 *
Dependents	1	1	125.31	<.0001 *
Dependents*Dependents	1	1	20.03	<.0001 *
Months_AFS	1	1	1.88	0.1704
Months_AFS_Bins	1	1	12.17	0.0005 *
Months_AFS*Months_AFS_Bins	1	1	34.01	<.0001 *
Unemp_Rate	1	1	8.90	0.0028 *
Unemp_Rate*Unemp_Rate	1	1	181.13	<.0001 *
Sex_Category*Pay_Grade	8	8	49.11	<.0001 *
Sex_Category*Marital_Stat	2	2	72.74	<.0001 *
Svc_Term*Dependents	6	6	28.65	<.0001 *
Svc_Term*Race	30	30	152.68	<.0001 *
Svc_Term*Education_Cat	24	24	82.12	<.0001 *
Reenl_Prohib*MOS_Branch	19	19	55.25	<.0001 *
Reenl_Prohib*Dependents	1	1	28.11	<.0001 *
Reenl_Prohib*Flags	2	2	60.49	<.0001 *
Reenl_Prohib*Accession_Type	5	5	212.94	<.0001 *
Reenl_Prohib*Months_AFS	1	1	75.56	<.0001 *
MOS_Branch*Marital_Stat	38	38	168.80	<.0001 *
MOS_Branch*Unemp_Rate	19	19	107.33	<.0001 *
Marital_Stat*Dependents	2	2	20.30	<.0001 *
Marital_Stat*Race	10	10	72.53	<.0001 *
Flags*Race	10	10	64.53	<.0001 *
Flags*Education_Cat	8	8	32.83	<.0001 *
Flags*Months_AFS	2	2	17.07	0.0002 *
Flags*Unemp_Rate	2	2	351.98	<.0001 *
Education_Cat*Unemp_Rate	4	4	16.39	0.0025 *
Accession_Type*Unemp_Rate	5	5	27.19	<.0001 *
Months_AFS*Unemp_Rate	1	1	162.29	<.0001 *

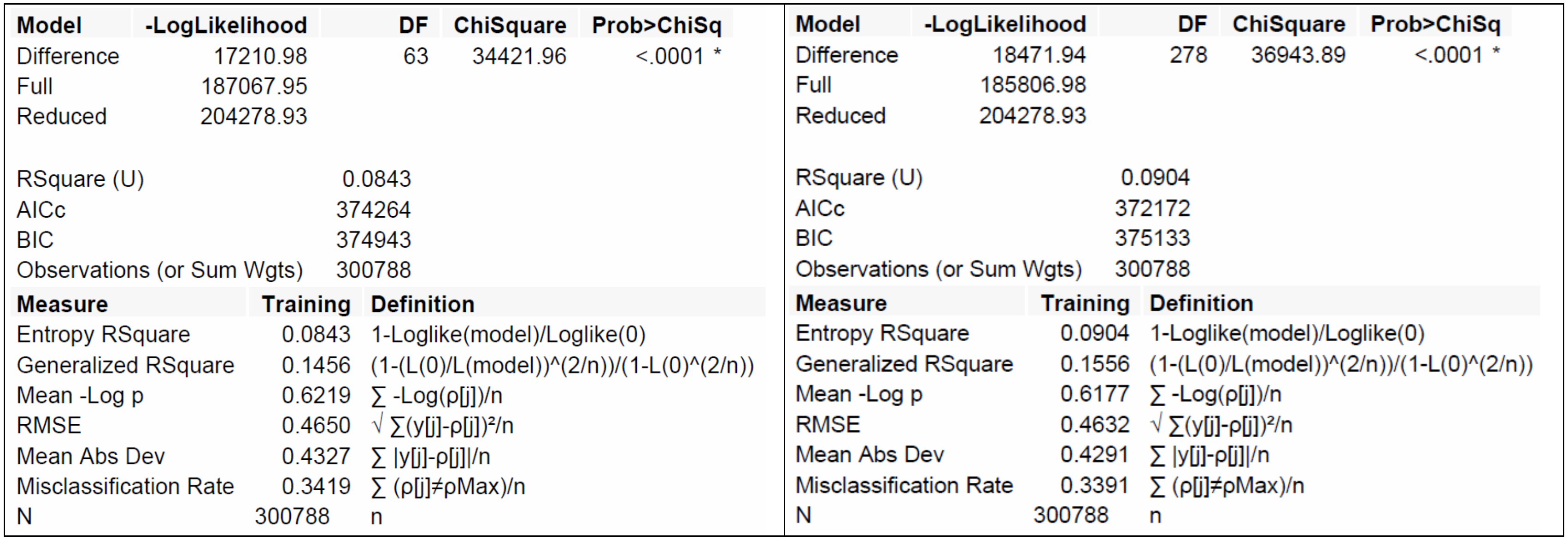
Table 12. Wald Tests on Model with Two-Way Interaction Terms

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ASSESSING THE FIT OF THE MODEL

A. SUMMARY MEASURES OF GOODNESS-OF-FIT

Table 13a shows several summary measures of the main effects model and Table 13b shows the same measures for the two-way interaction model. Higher is better for R-squared measures, and lower is better for the remaining measures such as residual mean squared error (RMSE) and misclassification rate. RMSE measures the average amount of squared error where error is defined as the difference between a soldier's actual outcome and the model estimate. The misclassification rate indicates the fraction of soldier outcomes that the model predicts incorrectly. As the tables indicate, the two-way interaction model performs slightly better on all of the summary measures.



Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	17210.98	63	34421.96	<.0001 *
Full	187067.95			
Reduced	204278.93			

RSquare (U)	0.0843
AICc	374264
BIC	374943
Observations (or Sum Wgts)	300788

Measure	Training	Definition
Entropy RSquare	0.0843	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1456	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6219	∑ -Log(ρ[j])/n
RMSE	0.4650	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.4327	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3419	∑ (ρ[j]≠ρMax)/n
N	300788	n

(a)

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	18471.94	278	36943.89	<.0001 *
Full	185806.98			
Reduced	204278.93			

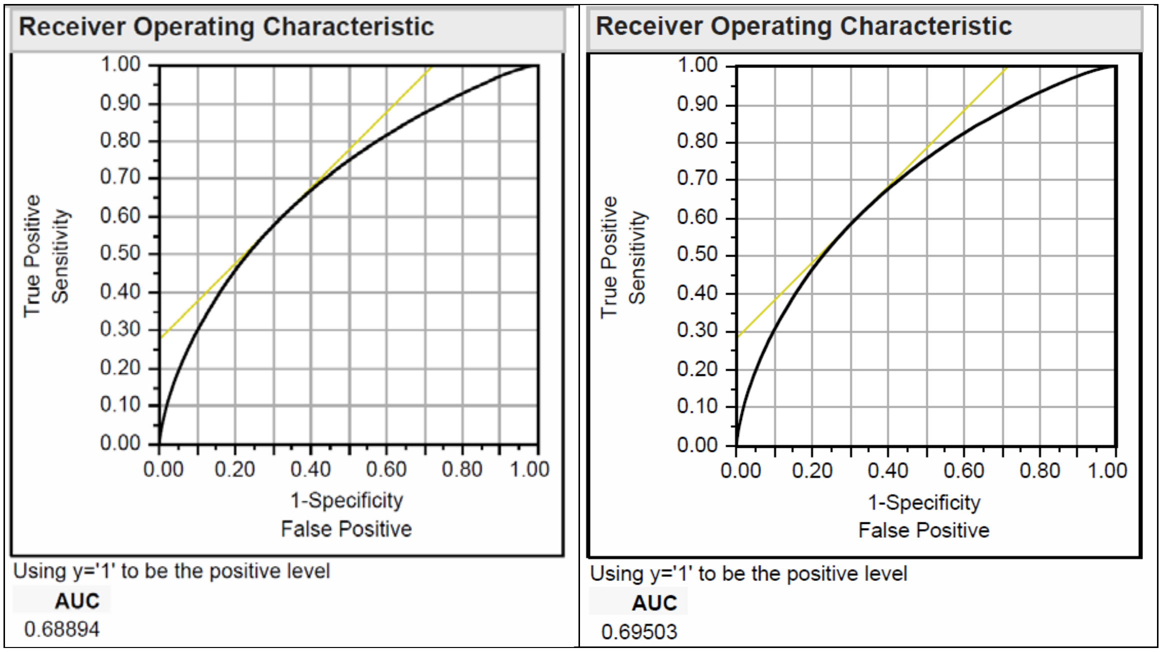
RSquare (U)	0.0904
AICc	372172
BIC	375133
Observations (or Sum Wgts)	300788

Measure	Training	Definition
Entropy RSquare	0.0904	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.1556	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6177	∑ -Log(ρ[j])/n
RMSE	0.4632	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.4291	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.3391	∑ (ρ[j]≠ρMax)/n
N	300788	n

(b)

Table 13. Summary Measures of (a) Main Effects Model, and (b) Two-Way Interaction Model

Yet another way to assess the goodness-of-fit of the models is by measuring the area under a receiver operating characteristic (ROC) curve. A ROC curve illustrates what will happen as the threshold for predicting a soldier will reenlist varies [10]. The most natural threshold is 0.5 so that the model will always select the most likely outcome. This threshold can be varied, however, in order to reduce the false positive rate or increase the true positive rate. When evaluating a ROC curve, higher area under the curve (AUC) is better. Figure 10 shows the ROC curve and AUC for the two models. As before, the two-way interaction model performs slightly better.



Using y='1' to be the positive level

AUC
0.68894

Using y='1' to be the positive level

AUC
0.69503

(a) (b)

Figure 10. ROC Curve and AUC for (a) Main Effects Model, and (b) Two-Way Interaction Model

B. CROSS-VALIDATION

Perhaps the best technique for determining the quality of models is cross-validation. Cross-validation involves splitting a data set into two parts, a training set and a test set. Then a model is generated using only the training data. This model must have the same covariates as the model built using the entire data set. Then predictions are made using the training model and the test set data, and the predictions are compared to the actual outcomes. In this study the training set consisted of the first 60 months of data, and the test set consisted of the last 7 months of data. Predictions were made by MOS and pay grade. Table 14 shows the main effect model predictions and actual outcomes for soldiers in the seven month test period.

There are many ways to assess the accuracy of model predictions. Residual sum of squares (RSS) sums the squared differences between the actual outcomes and the model estimates. It has the following form:

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

where y_i is the ith actual value, x_i is the ith covariate value, and $f(x_i)$ is the predicted value. RSS treats all the errors equally, regardless of the magnitude of the values involved. For example, if the model predicts 870 Infantry E-4 reenlistments and there are actually 869, then the error is 1. The same is true if the model predicts zero Armor E-9 reenlistments and there is actually one. Since the potential for large errors is much greater in categories where there are large numbers of soldiers involved, low RSS scores are usually due to accurately predicting the large groups of soldiers in a data set.

Mean absolute percentage error (MAPE) does just the opposite. MAPE scales the errors based on the size of the actual values involved. Low MAPE scores are usually due to accurately predicting the small groups of soldiers in a data set. It has the following form:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - f(x_i)}{y_i} \right|$$

where y_i is the ith actual value, x_i is the ith covariate value, and $f(x_i)$ is the predicted value.

Actual Reenlistments

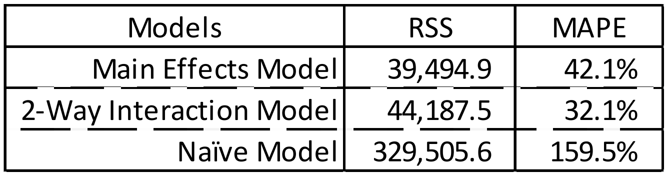
	E-1	E-2	E-3	E-4	E-5	E-6	E-7	E-8	E-9	Total
Adjutant General's Corps	1	5	25	365	183	126	27	1	0	733
Air Defense Artillery	2	2	14	142	64	23	0	1	0	248
Armor	1	5	23	275	160	67	4		0	535
Aviation	3	0	14	245	157	80	5		1	505
Chemical Corps	1	1	10	143	85	34	2	0		276
Civil Affairs/PsyOp	0			9	10	17	5	0	1	42
Combat Medic	3	1	27	354	222	98	16	1		722
Corps of Engineers	2	12	40	516	185	88	10		0	853
Field Artillery	0	9	40	413	217	97	14	1		791
Finance Corps	1		2	44	25	16	2	0	0	90
Infantry	3	14	70	869	513	346	28	3	2	1,848
Low Density	1	0	0	0					3	4
Medical Service Corps	0	2	7	172	126	54	2	0	0	363
Military Intelligence	1	3	16	198	210	186	30	1	0	645
Military Police Corps	0	4	17	303	185	93	14		0	616
Ordnance Corps	11	24	95	1,068	497	199	16	2	0	1,912
Quartermaster Corps	6	22	93	1,271	602	208	17	2	0	2,221
Signal Corps	3	6	40	575	362	145	19	0	0	1,150
Special Forces				8	11	147	55	1	0	222
Transportation Corps	1	6	37	497	281	93	6	1	0	922
Total	**40**	**116**	**570**	**7,467**	**4,095**	**2,117**	**272**	**14**	**7**	**14,698**

Predicted Reenlistments

	E-1	E-2	E-3	E-4	E-5	E-6	E-7	E-8	E-9	Total
Adjutant General's Corps	2.0	5.7	27.7	345.0	194.6	118.4	29.3	2.9	0.6	726.5
Air Defense Artillery	1.0	4.7	13.9	158.2	55.5	27.9	0.1	0.5	0.5	262.5
Armor	1.4	5.0	23.0	263.5	172.0	77.8	5.0		1.4	549.3
Aviation	1.0	1.5	14.6	262.1	184.7	71.0	6.4		0.9	542.2
Chemical Corps	1.0	3.1	11.1	133.6	92.1	33.2	1.9	0.2		276.2
Civil Affairs/PsyOp	0.1			7.4	23.3	16.7	3.5	0.6	0.4	52.0
Combat Medic	1.8	3.3	31.3	341.9	218.7	87.8	16.1	0.5		701.5
Corps of Engineers	4.7	13.9	50.9	469.3	164.3	85.3	10.1		0.3	798.8
Field Artillery	1.6	9.0	35.6	353.2	200.6	95.1	13.6	1.0		709.8
Finance Corps	0.5		4.3	47.1	17.1	12.4	1.6	0.1	0.3	83.4
Infantry	6.1	18.1	72.3	890.0	440.8	305.0	26.8	0.7	1.3	1,761.0
Low Density	12.4	0.1	0.6	0.4					1.5	15.1
Medical Service Corps	1.1	3.9	16.0	194.2	128.5	54.0	4.4	0.2	0.6	403.0
Military Intelligence	0.5	3.0	11.8	183.2	185.4	199.2	35.6	0.5	0.7	619.9
Military Police Corps	0.3	3.3	19.0	322.7	177.2	79.4	12.7		0.3	614.9
Ordnance Corps	7.1	27.5	108.3	1,060.1	496.8	180.2	17.2	2.0	1.0	1,900.2
Quartermaster Corps	8.6	34.2	139.5	1,374.9	591.3	197.9	16.6	3.9	1.5	2,368.4
Signal Corps	2.9	10.4	39.6	528.0	333.1	134.8	22.4	0.5	0.5	1,072.3
Special Forces				6.0	7.8	152.9	56.5	1.7	0.4	225.3
Transportation Corps	4.6	11.8	47.7	546.6	274.8	95.6	7.2	1.1	0.3	989.6
Total	**58.8**	**158.7**	**667.3**	**7,487.8**	**3,958.7**	**2,024.7**	**287.0**	**16.4**	**12.5**	**14,671.9**

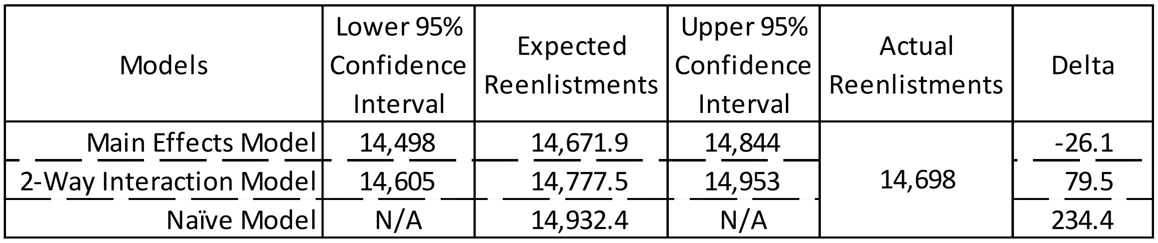
Table 14. Actual Reenlistments and Main Effects Model Predictions for Soldiers in the Seven Month Test Period

Tables 15 and 16 show the results of the main effects model, the two-way interaction model, and a naïve model. A naïve model assumes that all soldiers have a probability of reenlistment equal to the mean reenlistment rate of the soldiers in the training set. In this case the training set reenlistment rate was 41.73 percent. Fortunately both of the regression models perform much better than the naïve model. The two-way interaction model has the lowest MAPE, but the main effects model has the lowest RSS and predicts the total number of reenlistments more accurately. This means that the main effects model is the best option for predicting the large groups of soldiers, but the two-way interaction model is better at predicting small groups.



Models	RSS	MAPE
Main Effects Model	39,494.9	42.1%
2-Way Interaction Model	44,187.5	32.1%
Naïve Model	329,505.6	159.5%

Table 15. Cross Validation Results: RSS and MAPE



Models	Lower 95% Confidence Interval	Expected Reenlistments	Upper 95% Confidence Interval	Actual Reenlistments	Delta
Main Effects Model	14,498	14,671.9	14,844	14,698	-26.1
2-Way Interaction Model	14,605	14,777.5	14,953		79.5
Naïve Model	N/A	14,932.4	N/A		234.4

Table 16. Cross Validation Results: Expected Reenlistments

V. SUMMARY AND RECOMMENDATIONS

A. SUMMARY

The Army currently uses time series models to forecast personnel losses. Time series models can provide accurate predictions but offer no insights into the underlying causes of loss behavior. In order to quantify the various forces that influence retention rates, a regression model is necessary. A review of relevant literature reveals numerous potential covariates that may predict soldier retention rates. Most of these covariates were available in the data set and were included in two logistic regression models. The first model includes 14 main effects. The second model includes the same 14 main effects plus 21 highly significant two-way interaction terms. The two-way interaction model performs slightly better than the main effects model on all the summary measures, but the cross-validation results are mixed. Since the two-way interaction model is much more complicated to produce, and does not seem to generate results that are clearly better, the main effects model is probably the best option in most cases. Overall, both models estimate the total number of personnel that would reenlist over a seven-month test period fairly well. If a logistic regression model like the one described in this thesis was combined with an attrition model for soldiers who are not in their reenlistment window, it would be possible to make loss estimates for the entire active component enlisted force. These results could then serve as check on the current time series model estimates.

B. FOLLOW-UP RESEARCH RECOMMENDATIONS

1. Explore Additional Covariates

Several studies have focused on the impact of combat deployments on soldier retention. Most find that a limited number of deployments with adequate recovery time can actually boost retention rates, but frequent deployments with little recovery time usually reduce retention rates. In order to model this effect, one or more covariates must be carefully chosen. For instance, measuring the total number of combat deployments per soldier does not indicate the duration of the deployments, the amount of recovery time between them, or how long it has been since the soldier last deployed. A better measure might be the proportion of months deployed during the soldier's current term of service. This covariate would focus only on recent deployments, the proportion of months deployed rather than the number of deployments, and account for different enlistment term lengths. Interaction effects are also likely to be present. Soldiers in combat MOSs presumably have different experiences than soldiers in less dangerous support roles, and soldiers with families may be less willing to deploy repeatedly than single soldiers.

Another area that deserves more attention in the model is pecuniary factors. Since SRBs are the primary tool the Army uses to encourage reenlistments, it would be very useful to know the amount of money soldiers were eligible to receive for reenlisting. This would require a significant amount of historical research since bonus programs change all the time, and are often targeted only to soldiers that meet very specific criteria. Although information about soldiers that took SRBs must surely exist, the real challenge would be determining how much money soldiers declined.

The Rand study examining military and civilian pay [6] demonstrates that measuring changes in military compensation relative to civilian compensation would also be useful. This will be a difficult variable to quantify since compensation in the civilian workforce can vary dramatically by occupation, education level, location, and other factors. It may also to be difficult to quantify the value of healthcare benefits and pension income which only begin once a soldier gets out of the Army if they have at least 20 years of service.

2. Forecast Losses for the Entire Enlisted Force

In order to forecast losses for the entire enlisted force, additional logistic regression models are required in order to forecast losses for soldiers with fewer than 12 months until their ETS date, and an attrition model is necessary in order to forecast losses for soldiers with more than 12 months until their ETS date. This would be a very large project, but there is already an abundance of literature on attrition models so the researcher does not need to design either type of model from scratch. Once completed, the probability of retention could be estimated for every soldier over the next 12 months and then the results can easily be aggregated using a pivot table. The results could be used to check time series model results, provide insights into the causes of current loss behavior, and provide estimates of the impact of changing various Army retention policies on future retention rates.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX: REENLISTMENT PROHIBITION CODE DISTRIBUTION

Reenlistment Prohibition Codes	Description	Soldiers	% of Total
	Fully Eligible for Immediate Reenlistment	251,568	83.64%
10	Fully Eligible for Immediate Reenlistment	434	0.14%
11	Subject to Involuntary Separation	4,369	1.45%
9A	Lost Time	148	0.05%
9B	Adverse Action Flag	31	0.01%
9C	Denied Retention by SA – Commander Quality	44	0.01%
9D	Pending Security Clearance Determination	537	0.18%
9E	Physical Readiness	8,181	2.72%
9F	Denied Retention by Separation Authority	5	0.00%
9G	Grade (Soldier is within 24 months of ETS and exceeds RCP for Current Grade)	1,581	0.53%
9H	Pending MEB/PEB/MMRB	4,645	1.54%
9I	Non-Promotable Status	67	0.02%
9J	Involuntary Separation under Qualitative Service Program	2	0.00%
9K	Field Bar to Reenlistment	1,394	0.46%
9L	Involuntary Separation under Qualitative Management Program	6	0.00%
9M	Approved Retirement under Qualitative Management Program	23	0.01%
9N	Courts-Martial Conviction	38	0.01%
9O	Age (Restricted from Retention Due to Maximum Age Limitations)	1	0.00%
9P	Loss of Qualification in PMOS	474	0.16%
9Q	Declination of Continued Service Statement	3,963	1.32%
9S	Conscientious Objector	4	0.00%
9T	Approved Involuntary Separation	313	0.10%
9V	Pending Separation (Command or Soldier initiated separations)	956	0.32%
9W	Not Eligible Due to SSG NCOER/NCOES Eligibility Requirements	270	0.09%
9X	Other (Prohibitions not otherwise identified)	9,131	3.04%
9Y	Retirement (Application for retirement has been approved)	1,308	0.43%
9Z	Weight (Does not meet acceptable weight standards)	11,295	3.76%
		300,788	100.00%

Table 17. Reenlistment Prohibition Code Distribution

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] H. Golding and A. Adedeji, “Recruiting, retention, and future levels of military personnel,” United States Congressional Budget Office, Washington, DC, Oct. 2006.

[2] J. R. Hosek, J. Kavanagh, and L. Miller, “How deployments affect service members,” Rand, Santa Monica, CA, 2006.

[3] M. L. Hansen and S. Nataraj, “Expectations about civilian labor markets and Army officer retention,” Rand, Santa Monica, CA, 2011.

[4] B. J. Asch, P. Heaton, J. Hosek, F. Martorell, C. Simon, and J. T. Warner, “Cash incentives and military enlistment, attrition, and reenlistment,” Rand, Santa Monica, CA, 2010.

[5] M. C. Young, U. C. Kubisiak, P. J. Legree, and T. R. Tremble, "Understanding and managing the career continuance of enlisted soldiers," U.S. Army Research Institute for the Behavioral and Social Sciences, Fort Belvoir, VA, 2010, vol. 1280.

[6] J. R. Hosek, B. J. Asch, and M. G. Mattock, "Should the increase in military pay be slowed?" Rand Corporation, Santa Monica, CA, 2012.

[7] M. K. Brauner, K. L. Lawson, W. T. Mickelson, J. Adams, and J. M. Chaiken, “Time series models for predicting monthly losses of Air Force enlisted personnel,” Rand Corporation, Santa Monica, CA, 1991.

[8] A. Nelson, "Predicting enlisted reenlistment rates," M.S. thesis, Dept. Operations Research, Naval Postgraduate School, Monterey, CA, 2010.

[9] S. C. Orrick, "Forecasting Marine Corps enlisted losses," M.S. thesis, Dept. Operations Research, Naval Postgraduate School, 2008.

[10] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 2000.

[11] R. Powers. *About.com, Prior Service Enlistments* [Online]. Available: http://usmilitary.about.com/od/joiningthemilitary/a/priorservice.htm. Retrieved: 2013, Apr 28.

[12] United States Bureau of Labor Statistics. *Unemployment Rate – Civilian Labor Force* Series Id: LNS14000000 [Online]. Available: http://data.bls.gov/cgi-bin/surveymost?ln. Retrieved: 2012, Sep 23.

[13] TradingEconomics.com. *United States Consumer Confidence* [Online]. Available: http://www.tradingeconomics.com/united-states/consumer-confidence. Retrieved: 2012, Sep 23.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California